The Canadian Journal for the Scholarship of Teaching and Learning

Volume 6 | Issue 3 Article 2

12-11-2015

A Systematic Assessment of 'None of the Above' on Multiple Choice Tests in a First Year Psychology Classroom

Matthew V. Pachai
McMaster University, mattpachai@gmail.com

David DiBattista
Brock University, david.dibattista@brocku.ca

Joseph A. Kim
McMaster University, kimjoe@mcmaster.ca

A Systematic Assessment of 'None of the Above' on Multiple Choice Tests in a First Year Psychology Classroom

Abstract
Multiple choice writing guidelines are decidedly split on the use of 'none of the above' (NOTA), with some authors discouraging and others advocating its use. Moreover, empirical studies of NOTA have produced mixed results. Generally, these studies have utilized NOTA as either the correct response or a distractor and assessed its effect on difficulty and discrimination. In these studies, NOTA commonly yields increased difficulty when it is used as the correct response, and no change in discrimination regardless of usage. However, when NOTA is implemented as a distractor, rarely is consideration given to the distractor that could have been written in its place. Here, we systematically replaced each distractor in a series of questions with NOTA across different versions of an Introductory Psychology examination. This approach allowed us to quantify the quality of each distractor based on its relative discrimination index and assess the effect of NOTA relative to the quality of distractor it replaced. Moreover, our use of large Introductory Psychology examinations afforded highly stable difficulty and discrimination estimates. We found that NOTA increased question difficulty only when it was the correct response, with no effect on difficulty of replacing any distractor type with NOTA. Moreover, we found that NOTA decreased discrimination when it replaced the most effective distractors, with no effect on discrimination of replacing either the correct response or lowest quality distractor with NOTA. These results replicate the common finding that inclusion of NOTA as the correct response increases question difficulty by equally luring high-performing and low-performing students toward distractors. Moreover, we have shown that including NOTA as a distractor can reduce discrimination if used in lieu of a well written alternative, suggesting that multiple choice authors should avoid using NOTA on multiple choice tests.

Les guides de rédaction pour les questions à choix multiple sont incontestablement partagés sur l'usage de la réponse « aucune des situations ci-dessus ». Certains auteurs déconseillent de l'employer alors que d'autres préconisent de le faire. De plus, des études empiriques de l'emploi de cette expression ont mené à des résultats mitigés. En général, ces études ont utilisé l'option « aucune des situations ci-dessus » soit comme étant la réponse correcte soit comme distracteur et ont évalué ses effets sur la difficulté et le discernement. Dans ces études, la réponse « aucune des situations ci-dessus » mène généralement à une augmentation de la difficulté quand l'option est employée comme étant la réponse correcte et il n'y a aucun changement en ce qui concerne le discernement, quel que soit son usage. Toutefois, quand ce type d'option de réponse est utilisé en tant que distracteur, on trouve rarement une justification à l'emploi d'un distracteur qui aurait pu être utilisé à sa place. Dans le cas présent, nous avons systématiquement remplacé chaque distracteur dans une série de questions contenant l'option « aucune des situations ci-dessus » dans différentes versions d'un examen d'introduction à la psychologie. Cette approche nous a permis de quantifier la qualité de chaque distracteur sur la base de l'indice de son discernement relatif et d'évaluer les effets de l'option relative sur la qualité du distracteur qu'elle remplaçait. De plus, puisque nous avons utilisé de grands examens d'introduction à la psychologie, cela nous a permis de faire des estimations de la difficulté et du discernement hautement stables. Nous avons trouvé que l'emploi de l'option « aucune des situations ci-dessus » augmentait la difficulté de la question seulement lorsque cette option était la réponse correcte et qu'il n'avait aucun effet sur la difficulté présente lorsqu'on remplaçait n'importe quel type de distracteur par l'option « aucune des situations ci-dessus ». En outre, nous avons trouvé que l'emploi de l'option « aucune des situations ci-dessus » diminuait le discernement quand elle remplaçait les distracteurs les plus efficaces et qu'elle n'avait aucun effet sur le discernement quand elle remplaçait soit la réponse correcte soit le distracteur le moins plausible par l'option « aucune des situations ci-dessus ». Ces résultats reproduisent les conclusions communes selon lesquelles l'emploi de l'option « aucune des situations ci-dessus » comme étant la réponse correcte augmente la difficulté de la question car dans ce cas, tant les étudiants brillants que les étudiants médiocres sont leurrés de façon identique vers les distracteurs. Par surcroît, nous avons montré que le fait d'inclure l'option « aucune des situations ci-dessus » en tant que distracteur pouvait réduire le discernement si on l'utilisait à la place d'une alternative de réponse bien rédigée, ce qui suggère que les auteurs de questions à choix multiples devraient éviter d'utiliser l'option « aucune des situations ci-dessus » dans les examens à choix multiples.

Keywords
assessment, none of the above, multiple choice, discrimination

Multiple choice tests are a common form of assessment that most undergraduate students will complete during their education. A typical multiple choice question consists of a question stem and a series of options. Among this set of two to four distractor options is the key, which is coded as the correct response to the question. Given the ubiquity of multiple choice tests, it should not be surprising that a number of "best practice" guides for question writing have been produced (for a review see Haladyna, Downing, & Rodriguez, 2002). These guides generally consist of many recommendations, including posing the stem as a question or avoiding clues to the key. However, questions with significant flaws remain common on most multiple choice assessments (Downing, 2002, 2005; Jozefowicz et al., 2002). For example, in a sample of examinations distributed to first and second year medical students, Downing (2005) found flaws in 46% of the questions, with the number of flawed questions per test ranging from 36% to 65%. These flaws were violations of the recommendations made by Haladyna and colleagues (2002) and include unfocused stems, stems worded in the negative, and the use of an “all of the above” option. Flawed questions are concerning, as they can negatively impact student learning and disproportionately hinder the performance of the most knowledgeable students over others (Downing, 2005; Tarrant & Ware, 2008). Another concern regarding guides for multiple choice question writing is that the suggestions therein are infrequently based on empirical research (Frey et al., 2005). The goal of the present study was to address this concern in the context of one commonly used multiple choice option - none of the above (NOTA).

Multiple choice test writers are decidedly split in their opinions on the use of NOTA. In their review of test writing guidelines, Haladyna and colleagues (2002) noted that 48% of authors believe NOTA should never be used, while 44% believe NOTA has its place in multiple choice tests if implemented thoughtfully. Authors who advocate the use of NOTA cite its ability to increase question difficulty as a favourable outcome (Frary, 1991). One of the most common arguments against the use of NOTA is that it can reward students with no knowledge of the course content. Gross (1994) points out that when NOTA is the key, a student who does not know the true response can select NOTA and receive equal credit to a student who does know the true response. Gross argues any question format that readily rewards students with incorrect information, such as NOTA, should never be used.

Empirical research has also produced mixed results on the effect of NOTA. Most studies have found that NOTA increases test difficulty (Crehan & Haladyna, 1991; Dudycha & Carpenter, 1973; Forsyth & Spratt, 1980; Knowles & Welch, 1992; Oosterhof and Coats, 1984; Rimland, 1960; Tollefson, 1987; Wesman & Bennett, 1946). However, it is important to distinguish between the presence of NOTA as the key or as a distractor. In some studies demonstrating an effect of NOTA on item difficulty, the difference between NOTA as a key and NOTA as a distractor was not considered, and as a result these studies demonstrate relatively small effect sizes (Rimland, 1960). In other studies, NOTA increased difficulty when it was the key, but not when it was a distractor (Oosterhof & Coats, 1984; Tollefson, 1987). Another common measure of multiple choice item efficacy is the discrimination coefficient, which assesses how well a particular item distinguishes between high performing and low performing students as measured by their overall score on the test. Most studies demonstrate no effect of NOTA on discrimination, either as the key or as a distractor (Crehan & Haladyna, 1991; Dudycha & Carpenter, 1973; Tollefson, 1987). However, when students are asked to provide the correct response when choosing NOTA and their performance is rescored with regard to the accuracy of this response, including NOTA as a distractor does decrease discrimination coefficients (DiBattista, Sinnige-Egger, & Fortuna, 2014). Such a result suggests that previous work exploring the influence of NOTA on discrimination coefficients may have underestimated its true effect by pooling together NOTA-responders with and without accurate knowledge.

In the present study, we explored whether the effect of NOTA on item-wise difficulty and discrimination is modulated by the quality of distractor that could have been written in its place. Few empirical studies have evaluated the effect of NOTA relative to the other distractors, with most arbitrarily including NOTA in place of a random distractor (Dudycha & Carpenter, 1973) or as an additional distractor (Hughes & Trimble, 1965). In one case, NOTA replaced the best distractor, defined using a rank order of selection frequencies (Tollefson, 1987). This approach is flawed in that it assumes luring high performing and low performing students are equally desirable goals, when instead it may be preferable to lure lower-performing students to the distractors while higher-performing students select the key. In fact, the degree to which a particular question promotes this behavior is captured by the point-biserial correlation between accuracy for the given question and overall performance on the test, which is the most common measure of item-wise discrimination. However, this correlation provides no insight into the extent to which each distractor contributed to the question by luring lower-performing students, a true measure of the distractor's quality. For this purpose, we used the modified point-biserial correlation coefficient proposed by Attali and Fraenkel (2000). In this method, one creates a dichotomous variable capturing whether students select a given distractor or the correct response, and correlates it with overall test performance. This measure is functionally similar to the point-biserial discrimination coefficient, except that it quantifies the extent to which students answering incorrectly were lured by each distractor, respectively.

We designed this study with careful consideration to its implications for everyday teaching practice. Indeed, empirical studies of multiple-choice assessment have become increasingly important in light of recent evidence regarding the so-called Testing Effect, in which tests can enhance long-term learning, provided they are of sufficiently high quality (Roediger & Marsh, 2005). Moreover, our population (McMaster University's large Introductory Psychology course) afforded the sample size required to make meaningful manipulations and conclusions as well as the construct validity associated with a real classroom setting. On different versions of in-class midterm examinations, we manipulated the placement of NOTA such that it was not present, replaced the key, or replaced each of the three distractors for the same question across five different groups of students. This approach allowed us to measure the characteristics of our experimental questions in an independent group of students, then to measure how these characteristics change as a function of NOTA inclusion, specifically with regard to the quality of distractor NOTA replaced. Such a design models an instructor's decision to include NOTA or write an additional distractor in its place.

Method

Assessment in a Large Introductory Psychology Class

Our study was conducted using classroom tests deployed in large introductory psychology course at McMaster University. Each year, over 3000 students enroll in Introductory Psychology from faculties including Science, Social Science, Humanities, Business, and Nursing, making this course the largest on the McMaster University campus. To accommodate these enrollment levels, the course employs a unique blended learning approach to content delivery (Sana, Fenesi, & Kim, 2011). Most importantly, these enrollment levels allowed us to manipulate questions between students and be confident that our estimates of difficulty and discrimination were highly stable.

Conducting this study in a classroom afforded significant advantages over a laboratory study. In the laboratory, the test content must be general and appropriate for a random sample of participants and motivation to perform well is questionable. Conversely, a student’s performance on a classroom test is motivated by their actual academic goals, and the subject matter is intensive and varied. Topics covered in the Introductory Psychology course included research methods, learning, memory, cognition, social psychology, and personality theory. This study was conducted using the results from a multiple choice midterm examination that students in the course completed for 25% of their final grade.

Participants

Participants consisted of students completing their regularly scheduled midterm examination in Introductory Psychology over two academic years (2008/2009 and 2009/2010). Students consented to have their midterm examinations included in this study using a consent form that was included on the last page of their final examination, and a bonus mark was offered to students who responded, regardless of the nature of their response. This bonus encouraged students to consider the consent form before handing in their examination. We believe both lower-achieving students (who may have a greater need for grade incentives) and higher-achieving students (who may be generally motivated to maximize their grade) were incentivized equally to participate. All research protocols used were approved by McMaster University's Research Ethics Board. The total sample of consenting students was 1955 (63% participation) and 1742 (56% participation) on the two examinations.

Materials

On each midterm examination, a subset of the total questions was manipulated for inclusion in this study. The 2008/2009 examination included 5 experimental questions embedded in 25 total questions, and the 2009/2010 examination included 10 experimental questions embedded in 35 total questions, for an experimental *n* of 15 questions. On both examinations, the experimental questions were randomly distributed and were indistinguishable in style from non-experimental questions. NOTA never appeared in a non-experimental question, and all questions had one key and three distractors. See the Appendix for a sample of five experimental questions.

Design

Each examination consisted of five versions distributed randomly to the student population, where the five versions differed in the numerical placement of the questions. The non-experimental questions were identical on each test version, and the experimental questions were manipulated across test versions. For each experimental question, one test version contained the unmanipulated question (original), one version contained NOTA replacing the key, and the three remaining versions contained NOTA replacing each of the three distractors (see Table 1). When included, NOTA was always the final distractor, and the other distractors were shifted upward as necessary. This design ensured that, for each question, the measures of baseline question statistics measured in the *Original* condition were collected from a sample of students independent from those who completed the experimental conditions. Finally, to avoid systematic differences in difficulty across the five test versions, each test contained exactly three questions in each of the five experimental conditions.

Table 1
Example of Experimental Manipulation of Questions Across Test Versions

Question	Version 1	Version 2	Version 3	Version 4	Version 5
1	Original	replaced key	replaced D1	replaced D2	replaced D3
2	replaced D3	original	replaced key	replaced D1	replaced D2
3	replaced D2	replaced D3	original	replaced key	replaced D1
4	replaced D1	replaced D2	replaced D3	original	replaced key
5	replaced key	replaced D1	replaced D2	replaced D3	original

Distractor Analysis

Although there are many methods for classifying distractor quality, such as those provided by Item Response Theory, our methodological approaches were chosen for their ease of interpretation and availability to course instructors. Following completion of the examination, distractors were classified according to their discrimination ability in the O*riginal* condition. As recommended by Attali and Fraenkel (2000), distractor discrimination was quantified using $\mathrm{PB}_{\mathrm{DC}}$ (equation 1),

$$PB_{DC} = \frac{M_D - M_{DC}}{S_{DC}} \sqrt{\frac{P_D}{P_C}} \tag{1}$$

where M_{D} was the mean test score of students who selected the distractor, M_{DC} was the mean test score of students who selected either the distractor or the key, S_{DC} was the standard deviation in test scores of students who selected either the distractor or the key, P_{D} was the proportion of students selecting the distractor, and P_{C} was the proportion of students selecting the key. This measure quantifies the distinction between students who answered correctly and those who chose a given distractor, whereas the standard point-biserial discrimination coefficient quantifies the distinction between students who answered correctly and incorrectly without regard for which distractor they chose. It is important to note that $\mathrm{PB}_{\mathrm{DC}}$ was calculated using the data from students who answered the *Original* version of the question, which was always an independent sample of students from those who answered any of the NOTA manipulated versions. Using $\mathrm{PB}_{\mathrm{DC}}$, we ranked each distractor in a given question as the highest quality, medium quality, or lowest quality and measured the effect of replacing a distractor of a given quality with NOTA on item-wise difficulty and discrimination.

Results

Statistical analyses were conducted using R (R Development Core Team, 2015). Our dependent measures were difficulty and discrimination. Difficulty was defined as the percentage of students who answered a particular question correctly. Discrimination was defined as the point-biserial correlation between the responses for a particular question, either correct or incorrect, and the student population's total score on the examination. Each experimental question was considered to be a subject in a repeated-measures design. Mauchly’s test was used to detect violations of the sphericity assumption for repeated-measures designs. When sphericity was violated, the Greenhouse-Geisser correction, $\varepsilon_{\mathrm{GG}}$, was used to correct the resulting *p* values for univariate tests, and a multivariate ANOVA, which does not assume sphericity, was also computed for each dependent measure (Maxwell & Delaney, 2004). Effect size was expressed as Cohen's *f* for repeated-measures ANOVAs and Cohen’s *d* for paired-samples *t*-tests (Cohen, 1988).

Table 2 demonstrates difficulty, quantified by proportion correct, for each experimental question in each of the five conditions Although each question is considered a single subject in a repeated-measures design, it is important to note that each condition represents data collected from an independent sample of students, as questions were manipulated across versions of the test (see Table 1). These data are also represented in Figure 1. We analyzed these data using a one-way repeated-measures ANOVA, with NOTA placement as the within-subjects factor. Mauchly’s test revealed a violation of the sphericity assumption $W(9) = 0.16$, $p = 0.008$. A univariate Greenhouse-Geisser corrected ANOVA revealed a highly significant main effect of NOTA placement on question difficulty $F(4,56) = 12.99$, $\varepsilon_{GG} = 0.50$, $p < 0.0001$, $f = 0.80$. A multivariate ANOVA, which does not assume sphericity, also revealed a significant main effect (Wilks’ $A = 0.35$, $F(4,11) = 4.99$, $p = 0.0154$). To further analyze these differences, we performed post-hoc comparisons between the unmanipulated condition and each NOTA condition using two-tailed paired *t*-tests and a Bonferroni correction for multiple comparisons ($p_{\text{crit}} = 0.0125$). These comparisons revealed a significant effect of replacing the key with NOTA ($t(14) = 4.71$, $p = 0.0003$, $d = 1.22$), no significant effect of replacing the high quality ($t(14) = 0.098$, $p = 0.923$, $d = 0.03$) or medium quality ($t(14) = 1.20$, $p = 0.25$, $d = 0.31$) distractors with NOTA, and a nearly significant effect of replacing the low quality distractor with NOTA ($t(14) = 2.66$, $p = 0.019$, $d = 0.69$). Together, these results suggest that including NOTA as the key significantly increases question difficulty (lowers percent correct), while including NOTA as a distractor has little effect on difficulty.

Table 2
Difficulty (Proportion Correct) for Original and NOTA-Manipulated Questions

Question	Original	Key	High Quality	Med Quality	Low Quality
1	0.468	0.282	0.535	0.610	0.517
2	0.569	0.356	0.542	0.582	0.465
3	0.913	0.801	0.874	0.887	0.880
4	0.742	0.511	0.921	0.700	0.656
5	0.534	0.397	0.532	0.449	0.569
6	0.501	0.289	0.427	0.556	0.444
7	0.542	0.295	0.643	0.528	0.517
8	0.792	0.598	0.823	0.791	0.821
9	0.534	0.446	0.638	0.519	0.534
10	0.803	0.449	0.692	0.665	0.723
11	0.890	0.416	0.848	0.891	0.821
12	0.810	0.559	0.788	0.742	0.805
13	0.559	0.669	0.508	0.426	0.475
14	0.578	0.680	0.534	0.643	0.573
15	0.625	0.378	0.587	0.519	0.480
Mean	0.657	0.475	0.659	0.634	0.619
Median	0.578	0.446	0.638	0.610	0.569

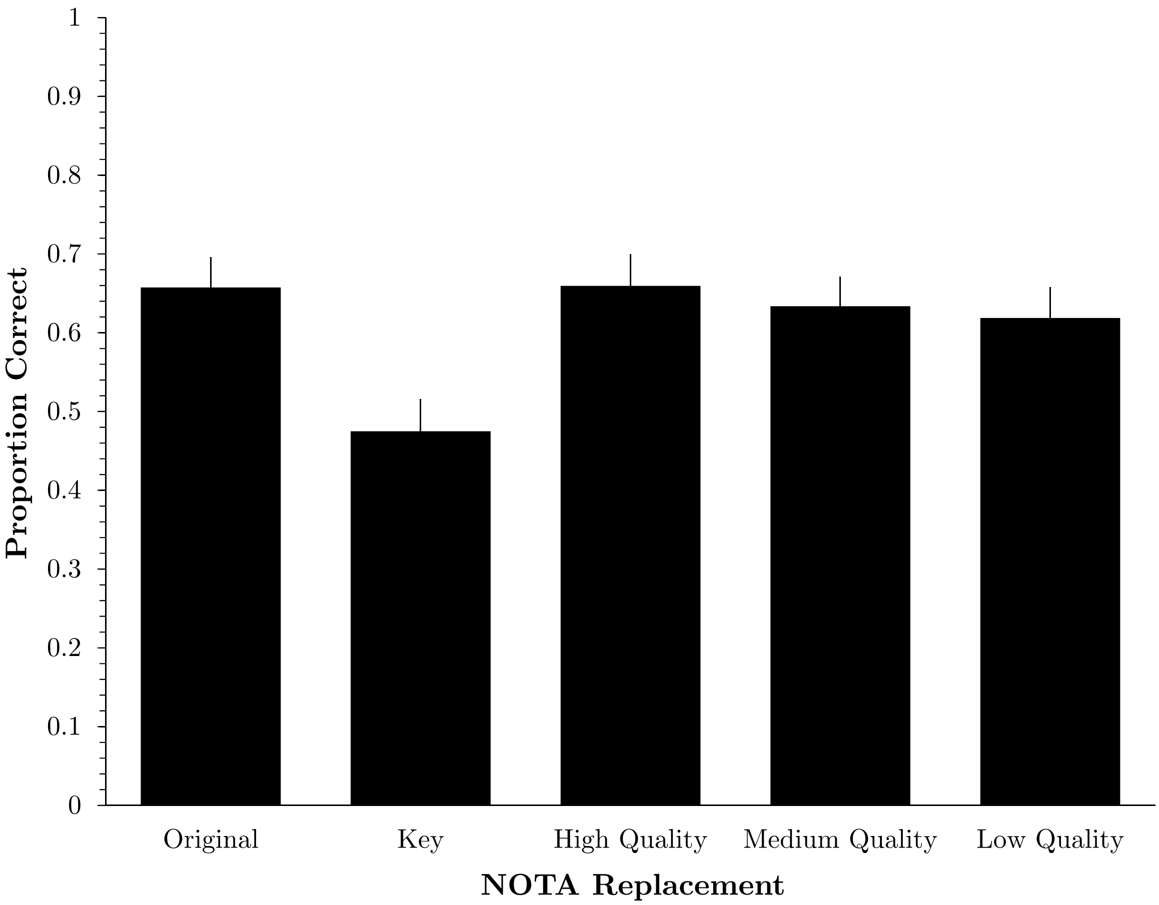

Figure 1. Item difficulty, measured as proportion correct, when NOTA was not present (Original), replaced the correct response (Key), or replaced a distractor (High, Medium or Low). Error bars represent +/- 1 SEM.

Table 3 demonstrates discrimination, quantified by the point-biserial correlation between accuracy and test-wise performance, for each experimental question in each of the five conditions. These data are also represented in figure 2. We analyzed these data using a one-way repeated-measures ANOVA, with NOTA placement as the within-subjects factor. Mauchly's test revealed a violation of the sphericity assumption $W(9) = 0.16$, $p = 0.008$. A Greenhouse-Geisser corrected univariate ANOVA revealed a main effect of NOTA placement that approached significance $F(4,56) = 2.34$, $\varepsilon_{GG} = 0.54$, $p = 0.11$, $f = 0.27$. A multivariate ANOVA, which does not assume sphericity, revealed a significant main effect of NOTA placement on discrimination (Wilks' $A = 0.31$, $F(4,11) = 6.13$, $p = 0.0076$). As with the difficulty results, we compared discrimination coefficients in the unmanipulated condition to each NOTA condition using two-tailed paired *t*-tests and a Bonferroni correction for multiple comparisons ($p_{\text{crit}} = 0.0125$). These comparisons revealed no significant effect of replacing the key with NOTA ($t(14) = 1.01$, $p = 0.32$, $d = 0.26$), a significant effect of replacing the high quality distractor with NOTA ($t(14) = 4.10$, $p = 0.001$, $d = 1.06$), a nearly significant effect of replacing the medium quality distractor with NOTA ($t(14) = 2.57$, $p = 0.02$, $d = 0.66$), and no significant effect of replacing the low quality distractor with NOTA ($t(14) = 1.28$, $p = 0.22$, $d = 0.33$). Together, these results suggest that including NOTA as the key has no effect on point-biserial correlation, but including NOTA as a distractor may decrease discrimination if it replaces a high-quality distractor.

Table 3

Discrimination (Point-Biserial Correlation) for Original and NOTA-Manipulated Questions

Question	Original	Key	High Quality	Med Quality	Low Quality
1	0.391	0.262	0.388	0.386	0.355
2	0.408	0.206	0.410	0.407	0.313
3	0.265	0.241	0.204	0.229	0.405
4	0.384	0.391	0.164	0.294	0.323
5	0.322	0.387	0.266	0.389	0.275
6	0.535	0.448	0.533	0.514	0.514
7	0.434	0.319	0.430	0.476	0.444
8	0.338	0.324	0.248	0.279	0.265
9	0.437	0.373	0.346	0.394	0.356
10	0.377	0.358	0.288	0.334	0.332
11	0.323	0.295	0.308	0.205	0.336
12	0.364	0.462	0.215	0.276	0.282
13	0.302	0.405	0.187	0.162	0.345
14	0.284	0.422	0.267	0.316	0.394
15	0.484	0.375	0.336	0.387	0.311
Mean	0.376	0.351	0.306	0.336	0.350
Median	0.377	0.373	0.288	0.334	0.336

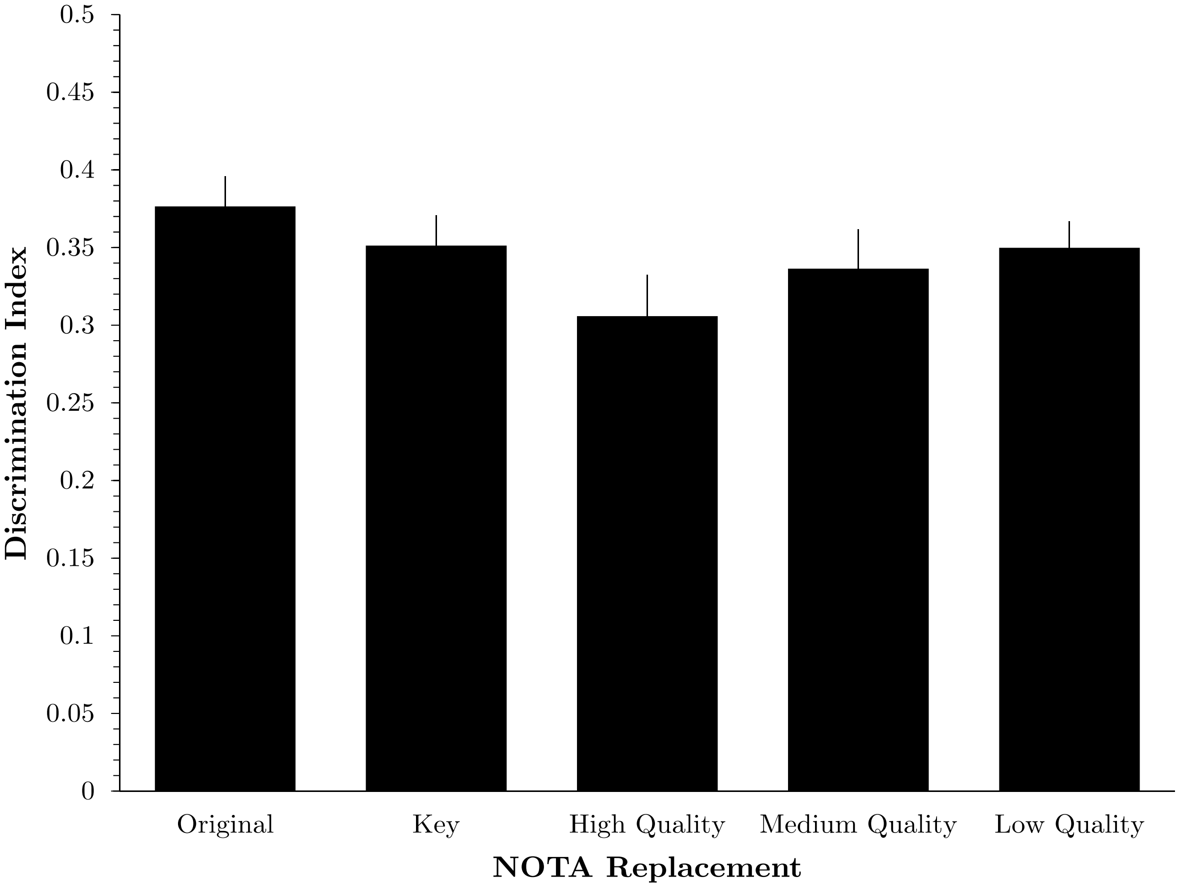

Figure 2. Item discrimination, measured as the point-biserial correlation between correct responses and test performance, when NOTA was not present (Original), replaced the correct response (Key), or replaced a distractor (High, Medium or Low). Error bars represent +/- 1 SEM.

Discussion

In the present study, we systematically manipulated the placement of the NOTA option on multiple choice questions and observed results that further elucidate the effect of NOTA in an educationally relevant setting. We found that NOTA significantly increased difficulty when it replaced the key. This finding is consistent with previous studies that have demonstrated an increase in difficulty with inclusion of NOTA (Crehan & Haladyna, 1991; Dudycha & Carpenter, 1973; Oosterhof & Coats, 1984; Tollefson, 1987). However, unlike all but a select few studies (Dudycha & Carpenter, 1973; Rich & Johanson, 1990), we have demonstrated a decrease in discrimination coefficients following NOTA inclusion. This effect was observed when NOTA replaced higher-quality distractors, suggesting that test-writers could obtain higher discrimination indices by including an additional high-quality distractor in lieu of NOTA. Indeed, when NOTA replaced higher quality distractors, discrimination suffered while difficulty was unchanged; in practical terms, this pattern suggests that fewer high-achieving students and more low-achieving students arrived at the correct response. We believe this to be an undesirable consequence of including NOTA, and consequently recommend that test writers avoid its inclusion on multiple-choice tests.

Our study was designed to permit practical recommendations for educational practice, but we must nonetheless qualify our results with their potential limitations. Administering experimental questions on classroom exams inherently risks alteration of typical test-taking behaviour. For this reason, Rich and Johanson (1990) suggested that NOTA should be used on no more than 25% of test items to prevent students from discounting its plausibility as a correct response. Our experimental items comprised 5 of 25 (20%) and 10 of 35 (28%) questions, respectively, for a total *N* of 15. Although our key result was quite robust, with a reduction of item-wise discrimination when NOTA replaced the highest-quality distractor in 14 of 15 items (see Table 3), this remains a relatively small number of experimental questions from which to draw general conclusions. Further research should explore the extent to which these results depend on the question stems or distractors employed, and elucidate the factors affecting the plausibility of NOTA as a distractor. We believe our results represent an important early step in such a research program, and an example of our experimental questions can be found in the appendix for the curious reader.

The present results, and those like them, have become increasingly important in light of a large body of literature suggesting that multiple-choice tests can also be used for the purposes of improving learning (Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Karpicke & Roediger, 2007; Roediger & Marsh, 2005; Thompson, Wenger, & Bartling, 1978). This observation, known as the Testing Effect, is derived from experiments demonstrating improvements in later recall when students have previously taken a test on some material as opposed to restudying it. However, multiple choice tests can also have negative consequences for learning (Roediger, 1996; McDermott, 2006; Roediger & Marsh, 2005). These negative consequences likely stem from what Remmers and Remmers (1926) termed the Negative Suggestion Effect, in which exposure to misinformation can increase the probability of recognizing or recalling that same misinformation on a later test. For example, Jacoby and Hollingshead (1990) found that exposure to misspelled words increased the probability of spelling errors on a later test, while Brown, Schilling, and Hockensmith (1999) found that the presentation of incorrect information between an initial test and a later test increased the probability of incorrect responses on the later test. For this reason, it is unsurprising that the benefits of multiple choice testing are strongest when students are able achieve a high level of performance (Roediger & Marsh, 2005; Butler & Roediger, 2008). Importantly, Odegard and Koen (2007) have demonstrated that when NOTA is used as the key, but not as a distractor, the positive testing effect is negated. They argue that when NOTA is the key, students can commit themselves to a distractor or select NOTA without knowing the true response, both of which lead to reinforcement of incorrect information. These results provide further reason to eliminate the NOTA option from multiple-choice tests.

Test-writers may include NOTA on multiple choice tests due to the challenge of writing numerous high-quality distractors for each question. However, a large body of literature has demonstrated the value of multiple choice questions with fewer options, since a higher number of such questions can be administered, permitting more thorough coverage of course content and higher test reliability with no cost to discrimination (Crehan, Haladyna, and Brewer, 1993; Haladyna & Downing, 1993; Owen & Froman, 1987; Trevisan, Sax, and Michael, 1991, 1994; see Rodriguez, 2005 for review). It may also be argued that including additional distractors reduces the impact of guessing. However, guessing should not be a significant concern for test developers because students with even a moderate level of engagement in the course material will rarely engage in truly random guessing, instead opting to eliminate distractors and select amongst the remaining responses (Ebel, 1968). For these reasons, we suggest that NOTA options should be replaced with higher-quality distractors, or more simply eliminated altogether.

Before making any decision regarding the construction of a test, instructors should carefully consider the goal of this assessment tool. On one hand, a test may be designed to challenge students who have not studied effectively, in which case inclusion of NOTA may be desirable for its effect on test difficulty. However, this same effect could be obtained, along with favourable effects on discrimination and reliability, by including higher-quality distractors or more questions with fewer distractors (Rodriguez, 2005). Moreover, decades of cognition research has suggested that testing can enhance student learning, and NOTA has been shown to counteract these positive effects (Odegard & Koen, 2007). Considered together, we believe the results of our study and those that come before it provide strong evidence against the inclusion of NOTA on any multiple choice test.

References

Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement*, *37*, 77–86. http://dx.doi.org/10.1111/j.1745-3984.2000.tb01077.x

Brown, A. S., Schilling, H. E. H., & Hockensmith, M. L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, *91*, 756–764. http://dx.doi.org/10.1037/0022-0663.91.4.756

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*, 604–616. http://dx.doi.org/10.3758/MC.36.3.604

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. http://dx.doi.org/10.3758/BF03202713

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Crehan, K. D., & Haladyna, T. M. (1991). The validity of two item-writing rules. *Journal of Experimental Education*, *59*(2), 183–192.

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*, 241–247. http://dx.doi.org/10.1177/0013164493053001027

DiBattista, D., Sinnige-Egger, J.-A., & Fortuna, G. (2014). The “none of the above” option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, *82*, 168–183. http://dx.doi.org/10.1080/00220973.2013.795127

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions : Do multiple-choice item-writing principles make any difference? *Medical Education*, *77*, 103–104.http://dx.doi.org/10.1097/00001888-200210001-00032

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Science Education*, *10*, 133–143. http://dx.doi.org/10.1007/s10459-004-4019-5

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*, 116–121. http://dx.doi.org/10.1037/h0035197

Ebel, R. L. (1968). Blind guessing on objective achievement tests. *Journal of Educational Measurement*, *5*, 321–325. http://dx.doi.org/10.1111/j.1745-3984.1968.tb00646.x

Forsyth, R. A., & Spratt, K. F. (1980). Measuring problem solving ability in mathematics with multiple-choice items: The effect of item format on selected item and test characteristics. *Journal of Educational Measurement*, *17*, 31–43. http://dx.doi.org/10.1111/j.1745-3984.1980.tb00812.x

Frary, R. B. (1991). The none-of-the-above option: An empirical study. *Applied Measurement in Education*, *4*, 115-124. http://dx.doi.org/10.1207/s15324818ame0402_2

Frey, B., Petersen, S., Edwards, L., Pedrotti, J., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*, 357–364. http://dx.doi.org/10.1016/j.tate.2005.01.008

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items: The case of “none of the above.” *Evaluation & the Health Professions*, *17*, 123–126. http://dx.doi.org/10.1177/016327879401700108

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*, 999–1010. http://dx.doi.org/10.1177/0013164493053004013

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309–334. http://dx.doi.org/10.1207/S15324818AME1503

Hogan, R., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567. http://dx.doi.org/10.1016/S0022-5371(71)80029-4

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple choice items. *Educational and Psychological Measurement*, *25*, 117–126. http://dx.doi.org/10.1177/001316446502500112

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology*, *44*, 345–358. http://dx.doi.org/10.1037/h0084259

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, *77*, 156–161. http://dx.doi.org/10.1097/00001888-200202000-00016

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. http://dx.doi.org/10.1016/j.jml.2006.09.004

Knowles, S. L., & Welch, C. A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using "none-of-the-above." *Educational and Psychological Measurement*, *52*, 571–577. http://dx.doi.org/10.1177/0013164492052003006

Maxwell, S., & Delaney, H. (2004). *Designing experiments and analyzing data: A model comparison approach* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

McDermott, K. B. (2006). Paradoxical effects of testing: repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*, 261–267. http://dx.doi.org/10.3758/BF03193404

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: implications for the testing effect. *Memory*, *15*, 873–885. http://dx.doi.org/10.1080/09658210701746621

Oosterhof, A. C., & Coats, P. K. (1984). Comparison of difficulties and reliabilities of quantitative word problems in completion and multiple-choice item formats. *Applied Psychological Measurement*, *8*, 287–294. http://dx.doi.org/10.1177/014662168400800305

Owen, S. V., & Froman, R. D. (1987). What's wrong with three-option multiple choice items? *Educational and Psychological Measurement*, *47*, 513–522. http://dx.doi.org/10.1177/0013164487472027

R Development Core Team. (2015). *R: A language and environment for statistical computing.* Vienna, Austria.

Remmers, H., & Remmers, E. M. (1926). The negative suggestion effect of true false examination questions. *The Journal of Educational Psychology*, *17*, 52–56. http://dx.doi.org/10.1016/j.athoracsur.2010.01.082

Rich, C. E., & Johanson, G. A. (1990). *An item-level analysis of "none of the above".* Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Rimland, B. (1960). The effects of varying time limits and of using "right answer not given" in experimental forms of the U.S. Navy arithmetic test. *Educational and Psychological Measurement*, *20*, 533–539. http://dx.doi.org/10.1177/001316446002000310

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items : A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*, 3–13. http://dx.doi.org/10.1111/j.1745-3992.2005.00006.x

Roediger, H. L. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language*, *35*, 300–318. http://dx.doi.org/10.1006/jmla.1996.0017

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159. http://dx.doi.org/10.1037/0278-7393.31.5.1155

Sana, F., Fenesi, B., & Kim, J. A. (2011). A case study of the introductory psychology blended learning model at McMaster University. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(1). http://dx.doi.org/10.5206/cjsotl-rcacea.2011.1.6

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42, 198-206. http://dx.doi.org/10.1111/j.1365-2923.2007.02957.x

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning & Memory*, *4*, 210–221. http://dx.doi.org/10.1037/0278-7393.4.3.210

Tollefson, N. (1987). A comparison of the item difficulty and item discrimination of multiple-choice items using the "none of the above" and one correct response options. *Educational and Psychological Measurement*, *47*, 377–383. http://dx.doi.org/10.1177/0013164487472010

Trevisan, M. S., Sax, G., & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement, 51*, 829–837. http://dx.doi.org/10.1177/001316449105100404

Trevisan, M. S., Sax, G., & Michael, W. B. (1994). Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54*, 86–91. http://dx.doi.org/10.1177/0013164494054001008

Wesman, A. G., & Bennett, G. K. (1946). The use of "none of these" as an option in test construction. *Journal of Educational Psychology, 37*, 541-549. http://dx.doi.org/10.1037/h0056815

APPENDIX

Sample multiple-choice questions, presented in their original format. NOTA was added to each question across test versions as described in Table 1. Correct responses are underlined.

1. **Harlow's** studies of infant monkeys raised with **surrogate** mothers indicated that infants became attached to the surrogate mother:

   a) from which **food** was most often delivered.
   b) that provided the most **contact comfort.**
   c) that was present when **danger** was presented.
   d) that was present for the greatest amount of **time**.

2. The terminal **bouton** of Neuron A forms a **synapse** with the cell body of Neuron B. Neuron A begins firing action potentials, releasing neurotransmitters that bind to Neuron B, causing **EPSPs**. However, Neuron B has **not** fired an action potential. What may be causing this?

   a) Sodium ions entering the dendrites of a neuron can cause an action potential, but sodium ions entering the cell body **cannot**.
   b) The sodium ions may not hold a strong enough positive **charge** to depolarize the resting potential of Neuron B.
   c) Neuron B's **chlorine** channels are being opened at the same time by a third neuron.
   d) Neuron B's **sodium-potassium pump** is active, ejecting sodium ions from the cell

3. An experimenter asks you to complete two visual search tasks. In the first, you search for a **green square** in an array of **red squares**. In the second, you search for a **green square** in an array of **green circles** and **red squares**. As set size **increases**, what should happen to your **response time** in the first experiment? In the second experiment?

   a) Increase in both experiments
   b) No change in the first, increase in the second
   c) Increase in the first, decrease in the second
   d) No change in the first, decrease in the second

4. Damage to the **cerebellum** would most likely result in:

   a) uncontrollable bursts of panic
   b) difficulty sleeping or staying awake
   c) jerky, uncoordinated movements.
   d) numbness on one side of the body.

5. Your elderly neighbour Violet is a subject in a study comparing the ability of a new drug to relieve arthritis pain. The drug will be administered to Violet by one of the physicians conducting the study. If it is a **double-blind, within-subjects** design then we would expect that while participating in the study Violet will:

    a) receive **either** a drug or a placebo, and **only** the physician will know which she is getting.
    b) at **different times**, receive **both** the drug and the placebo, but **neither** she nor the physician will know which she is getting.
    c) receive **either** a drug or a placebo, and **neither** she nor the physician will know which she is getting.
    d) **at different times**, receive **both** the drug and the placebo, but **only** the physician will know which she is getting